Language and Cognition **12** (2020), 385–410. doi:10.1017/langcog.2020.5

Teaching the unlearnable: a training study of complex *yes/no* questions*

BEN AMBRIDGE

University of Liverpool, ESRC International Centre for Language and Communicative Development (LuCiD)

CAROLINE F. ROWLAND

Max Planck Institute for Psycholinguistics, University of Liverpool, ESRC International Centre for Language and Communicative Development (LuCiD), Donders Institute for Brain, Cognition and Behaviour, Radboud University

AND

ALISON GUMMERY

University of Liverpool, ESRC International Centre for Language and Communicative Development (LuCiD)

(*Received* 14 *May* 2019 – *Revised* 3 *January* 2020 – *Accepted* 3 *January* 2020)

ABSTRACT

A central question in language acquisition is how children master sentence types that they have seldom, if ever, heard. Here we report the findings of a pre-registered, randomised, single-blind intervention study designed to test the prediction that, for one such sentence type, complex questions (e.g., *Is the crocodile who's hot eating?*), children could combine schemas learned, on the basis of the input, for complex noun phrases (*the [THING] who's [PROPERTY]*) and simple questions (*Is [THING] [ACTION]ing?*) to yield a complex-question schema (*Is [the [THING] who's [PROPERTY]] ACTIONing?*). Children aged 4;2 to 6;8 (M = 5;6, SD = 7.7 months) were trained on simple questions (e.g., *Is the bird cleaning?*) and either (Experimental group, N = 61) complex noun phrases (e.g., *the bird who's sad*) or (Control group, N = 61) matched simple noun

[*] This work was supported by the International Centre for Language and Communicative Development (LuCiD).The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. Address for correspondence: tel: +44 151 794 1111; e-mail: Ben.Ambridge@Liverpool.ac.uk

phrases (e.g., *the sad bird*). In general, the two groups did not differ on their ability to produce novel complex questions at test. However, the Experimental group did show (a) some evidence of generalising a particular complex NP schema (*the [THING] who's [PROPERTY]* as opposed to *the [THING] that's [PROPERTY]*) from training to test, (b) a lower rate of auxiliary-doubling errors (e.g., **Is the crocodile who's hot is eating?*), and (c) a greater ability to produce complex questions on the first test trial. We end by suggesting some different methods – specifically artificial language learning and syntactic priming – that could potentially be used to better test the present account.

KEYWORDS: complex syntax, complex questions, structure dependence, yes/no questions, training study, language acquisition

1. Introduction

Few questions are more central to our understanding of cognitive development than that of how children learn to produce sentences in their native language. For sentences that children have frequently heard in exactly that form (e.g., the question *What's that?*; typically one of the first produced by learners of English; e.g., Ambridge, Rowland, Theakston, & Tomasello, 2006), the answer is potentially straightforward: children repeatedly hear the form paired with a particular inferred meaning (in this case, something like 'I request that you produce the conventional label for this object'), and so produce this form when they want to convey this meaning themselves. For sentences that children have not heard (e.g., *The hungry mouse chased the shy panda*) the problem is considerably more difficult, but it is still tractable. Children hear very many sentences of the same type (e.g., *The old man bought a new book; The little girl ate a delicious cake*), and generalise based on the shared properties of basic two-participant sentences (e.g., shared sentence structure, meaning, syntactic or semantic roles); though the precise nature of this generalisation varies considerably from theory to theory (e.g., Pinker, 1984; Wexler, 1998; Tomasello, 2003; Sakas & Fodor, 2012).

But, for a third type of sentence, the problem appears completely intractable. Consider, for example, the sentence *Is the crocodile who's hot eating?* (an example from the present study). Not only have children never heard this PARTICULAR sentence (a Google search yields no results), they will have rarely, if ever, heard ANY sentence of this form (i.e., a *yes/no* question that contains a relative clause [e.g., *who's hot*]). In a search of approximately three million child-directed utterances in the CHILDES database, MacWhinney (2004) found only a single example (cf., Cowie, 1998; Pullum & Scholz, 2002). Yet, although children struggle with these complex-questions (showing around only 65% accuracy at age six to seven; Ambridge, Rowland, & Pine, 2008), at some point in the transition to adulthood, they master them.

How, then, do children "learn the unlearnable" (Regier & Gahl, 2004, p. 147)? Historically (e.g., Chomsky, 1980; Crain & Nakayama, 1987), a popular answer has been that they cannot. Children cannot learn the correct structure of complex *yes/no* questions solely from the input, and master it only with the aid of two pieces of innate knowledge (i.e., knowledge with which learners are born). The first is knowledge that some languages use a movement rule to transform statements into questions by moving the auxiliary (here, *is*) (1):

(1) the crocodile **is** eating → **is** the crocodile ~~**is**~~ eating?

The second is knowledge that this rule is STRUCTURE DEPENDENT (i.e., is formulated in terms of the structural constituents of the sentence, as opposed to, say, the linear ordering of the words). For complex questions in English, the correct structure-dependent rule is 'move the auxiliary IN THE MAIN CLAUSE' (2):

(2) the crocodile who's hot **is** eating → **is** the crocodile who's hot ~~**is**~~ eating?

and not (for example) 'move the first auxiliary' (3):

(3) the crocodile who**'s** hot is eating → ***is** the crocodile who~~**'s**~~ hot is eating?

Indeed, the ability of children to produce complex questions without ever hearing them is often taken as the "parade case" (Crain, 1991, p. 602) for innate knowledge of language (see also Crain & Pietroski, 2001; Laurence & Margolis, 2001; Fodor & Crowther, 2002; Legate & Yang, 2002; Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2008; Berwick, Pietroski, Yankama, & Chomsky, 2011; other frequently discussed cases include the *wanna* contraction, e.g., Crain & Thornton, 1998; Getz, 2019; and anaphoric *one*, e.g., Lidz, Waxman, & Freedman, 2003; Akhtar, Callanan, Pullum, & Scholz, 2004; Pearl & Mis, 2016; Goldberg & Michaelis, 2017. For reviews, see Pullum & Scholz, 2002; Scholz & Pullum, 2002, 2006).

Our goal in the present study was to investigate a different possibility: that, although children never hear complex questions, they learn to produce them by combining constructions that they have learned from the input (Stemmer, 1981; Sampson, 1989; Ambridge, Rowland, & Pine, 2008; Ambridge & Rowland, 2009; Clark & Lappin, 2011; Fitz & Chang, 2017; see also Lewis & Elman, 2001; Reali & Christiansen, 2005; Perfors, Tenenbaum, & Regier, 2011, for alternative learning approaches). To lead with our conclusion, we were largely unsuccessful in our attempt to test this theoretical proposal, largely because of the fact that many of the children studied had already mastered complex *yes/no* questions, rendering our experimental manipulation moot. Nevertheless, we feel that the paper makes two important methodological contributions by (a) setting out a training methodology that could profitably be used to investigate acquisition of other constructions, and (b) leading us closer to an understanding of what

methods likely will and will not work for future investigations of the acquisition of complex questions (we compare three possible approaches in the 'Discussion').

The schema-combination proposal that we set out to test is, intuitively, a very simple one. Children hear thousands of simple *yes/no* questions, and SCHEMATISE across them (Tomasello, 2003) to form a **slot-and-frame pattern**, which allows them to produce new questions by inserting new items into the slots (e.g., Dąbrowska, 2000; Dąbrowska & Lieven, 2005; Rowland, 2007):

Is the bird cleaning?
Is the fish swimming?
Is the whale falling?
Is the frog clapping?

Is [THING] [ACTION]ing?

Similarly, children hear thousands of complex noun phrases, and schematise across them to produce a slot-and-frame pattern which allows for the production of new exemplars:

the cow who's small
the bird who's big
the dog who's white
the mouse who's black

the [THING] who's [PROPERTY]

The final, crucial, step is SCHEMA-COMBINATION: The complex-noun-phrase schema ***the [THING] who's [PROPERTY]*** itself denotes a THING, and so can be inserted into the [THING] slot of the *yes/no* question schema:

Is [the [THING] who's [PROPERTY]] ACTIONing?

Unlike complex questions, such forms can be found in child-directed speech. An automated search of the child-directed-speech portion of the Manchester corpus, consisting of 12 children aged two to three years (Theakston, Lieven, Pine, & Rowland, 2001), yielded roughly 1,300 questions of the form *Is X Ying?*, around 100 of which were of the form *Is* ***the*** *X Ying?*. (Note further than even these 1,300 are only a relatively small subset of the broader class of *yes/no* questions; e.g., *Are you … ? Do you … ? Does it … ?*). Complex noun phrases were rarer, but by no means absent, even in this relatively sparsely sampled corpus. For example, a search for the phrase *the X that's / that is Ying* yielded 13 examples:

and this is **the crane** isn't it **that's lifting the blocks**
oh is that **the shark that's going** to eat dolly up

there's **the one that is occasioning** you some difficulty
where's **the security man that's going** to deliver this
don't think that's **the bunny+rabbit that's missing** though
as_well_as **the one that's lying down**
you're **the one that's being** naughty
you're **the one that's pulling** it apart
it's you **the one that's leaning** on me
I don't think it's **the baby tiger that's being** naughty
what color's **the piece that's sticking** out
what about **the other piece that's sticking** out
the bit that's burning hot

A search for the phrase *the X who's / who is Ying* yielded a further four examples:

well if it rolls over the man **the man who's sitting** driving it might get crushed
you're **the one who's being** silly
because you're **the one who's giving** everybody colds
that's **the medicine I think for that doctor who's looking** after that animal

Admittedly, the number of complex noun phrases of the exact form *the X that['s/ is] / who['s/ is] Ying* (17) is not large. However, if we assume that learners perceive at least some similarity between complex noun phrases with different forms (e.g., between *the one that's lying down*, *the one that can lie down*, *the one that you like*, etc.), this number increases dramatically. A detailed count cannot easily be done automatically, and so is beyond the scope of this paper; but MacWhinney (2004, p. 890) notes that, even if – purely for ease of identification – we narrow the search to complex noun phrases that happen to occur in *wh-* questions, "there are hundreds of input sentences of this type in the CHILDES corpus … [for example] *Where is **the dog that you like?***)". In essence, our proposal follows MacWhinney (2004, p. 891) in assuming that children use the evidence available in the input to "learn to fill argument slots with compound constituents" (see also MacWhinney, 1975, p. 1987).

Note that, while the availability in the input of simple questions and complex noun phrases is good news for the prima facie plausibility of the schema-combination proposal, it is bad news for the present training study: ideally, children would have had very little opportunity to learn the structure of complex *yes/no* questions prior to the study, allowing the effects of training (in the Experimental but not Control group) to shine through. In fact, the above corpus analysis suggests that the children tested have already had ample opportunity to learn the structure of complex *yes/no* questions, and that the relatively small amount of additional relevant input provided during the training study is likely to have limited impact. Indeed, this proved to be the case.

Surprisingly, given its centrality to the issue of language learnability, and the availability of relevant input, the question of whether children can learn to produce complex questions by schema combination has never been tested (though see Abbot-Smith & Behrens, 2006, for evidence of children's ability of schema-combination with regard to German passive and future forms). In the present paper, we report a pre-registered, randomised, single-blind intervention study that was designed to answer the question. An Experimental group were given training on simple *yes/no* questions and complex noun phrases, as in the examples above. A Control group (matched for language ability) were given training on simple *yes/no* questions and – instead of complex noun phrases – semantically matched simple adjectival noun phrases (e.g., *the small cow* and the *big bird* rather than *the cow who's small* and *the bird who's big*). Our pre-registered hypothesis was that the Experimental group would outperform the Control group on both (a) the number of correct complex questions produced and (b) the number of children able to produce at least one correct complex question.

2. Method

2.1. ETHICS AND PRE-REGISTRATION

This study was approved by the University of Liverpool Ethics Committee. The methods, hypotheses, sampling, and analysis plan were registered at the website of the Open Science Framework prior to the collection of any data. All training and test materials, analysis scripts and anonymised raw data, can be found at <https://osf.io/e2q54/; <doi:10.17605/OSF.IO/E2Q54>.

2.2. PARTICIPANTS

A pre-registered sample size of 122 children (61 per group, randomly allocated) was determined on the basis of a power calculation with $d = 0.3$, *power* = 0.5, on the basis of a between-subjects *t*-test (using GPower 3.0). An effect size of $d = 0.3$ was chosen not on the basis of previous work, since we were not able to identify any similar studies, but solely with reference to the rules of thumb set out in Cohen's (1992) *Power Primer*. Somewhat optimistically, we chose an effect size slightly larger than that designated 'Small' for a test for independent means ($d = 0.2$), but smaller than that designated 'Medium' ($d = 0.5$). Although our analysis plan actually specified the use of mixed-effects models, it is not possible to run a power analysis for such models without simulated data, and we were not aware of any findings from studies with sufficiently similar methods to form the basis for such a simulation. Although a power greater than 0.5 would have been desirable, a total sample size of 122 was our maximum in terms of time, funding, and personnel. We go some way towards

mitigating this problem by also including a supplementary, exploratory Bayesian analysis (the decision to add this analysis was taken after the main results were known). A total of 143 children completed the experiment, but 21 were excluded (9 from the Experimental group and 12 from the Control group) for failing to meet the pre-registered training criteria set out below. Children were recruited from UK Reception (aged four to five years) and Year 1 (five to six years) classes. The final sample ranged from 4;2 to 6;8 (M = 5;6, SD = 7.7 months; Experimental group: M = 64.85 months, SD = 7.93; Control group: M = 66.54 months, SD = 7.44).

2.3. STANDARDISED TEST

Before training, all participants completed the Word Structure test from the fifth edition of the CELF-Preschool 2 UK (Wiig, Secord, & Semel, 2004). This is a production test of morphosyntax, in which children are asked to complete sentences to describe pictures (e.g., Experimenter: "This girl is climbing. This girl is …" Child: "Sleeping"). The purpose of this test was to allow us to verify that the Experimental and Control groups were matched for general ability with morphosyntax. This was found to be the case (Experimental: M = 19.42, SD = 3.05; Control: M = 19.95, SD = 2.79). We did not include a baseline measure of complex-question production because we did not want to give children practice in producing these questions, since our goal was to investigate the impact of relevant training on children who had previously heard no – or extremely few – complex questions. In retrospect, this decision was unfortunate, since it left us unable to confirm a conclusion suggested by our findings: that a large proportion of children already had at least some knowledge – and some a firm grasp – of the structure of complex *yes/no* questions before the study began.

2.4. TRAINING

All participants completed five training sessions on different days. As far as possible, children were tested on five consecutive days, but sometimes this was not possible due to absence. The total span of training (in days) for each child was included as a covariate in the statistical analysis. Each daily training session comprised two sub-sessions: Noun Phrases and simple *yes/no* questions, always in that order. The CELF was presented immediately before the first training session on Day 1; the complex-question test session immediately after the final training session on Day 5.

2.4.1. *Noun-Phrase (NP) training*

The aim of this part of the session was to train children in the Experimental group on complex noun phrases (e.g., *the bird who's happy*), resulting in the

TABLE 1. *Day 1 Noun-phrase training for children in the Experimental and Control groups. Children heard each NP in the 'Experimenter' column, and heard and repeated each NP in the 'Child' column*

Experimental group		Control group	
Experimenter	Child	Experimenter	Child
the bird who's happy	the bird who's sad	the happy bird	the sad bird
the fish who's happy	the fish who's sad	the happy fish	the sad fish
the whale who's sad	the whale who's happy	the sad whale	the happy whale
the frog who's sad	the frog who's happy	the sad frog	the happy frog
the chicken who's big	the chicken who's small	the big chicken	the small chicken
the cow who's big	the cow who's small	the big cow	the small cow
the bear who's small	the bear who's big	the small bear	the big bear
the horse who's small	the horse who's big	the small horse	the big horse
the dog who's black	the dog who's white	the black dog	the white dog
the cat who's black	the cat who's white	the black cat	the white cat
the mouse who's white	the mouse who's black	the white mouse	the black mouse
the rabbit who's white	the rabbit who's black	the white rabbit	the black rabbit

formation of a complex-noun-phrase schema (**the [THING] who's [PROPERTY]**) that could be combined with a simple question schema (**Is [THING] [ACTION]ing?**) to yield a complex-question schema (**Is [the [THING] who's [PROPERTY]] ACTIONing?**). (In retrospect, it would have been preferable to use *that* instead of *who*, since – as we discovered only after running the study – the former is more frequent in the input and appears to be children's preferred form, at least for animals). On each day, children in the Experimental group heard the experimenter produce 12 such complex noun phrases (see Table 1, first column), and heard and repeated a further 12 such phrases (see Table 1, second column).

NP training took the form of a bingo game, in which the experimenter and child took turns to request cards from a talking dog toy, in order to complete their bingo grid. A similar method has been used to elicit sentences in syntactic priming studies (e.g., Rowland, Chang, Ambridge, Pine, & Lieven, 2012; Peter, Chang, Pine, Blything, & Rowland, 2015), but the present study is novel in adapting this method for use in a training study, as well as for eliciting NPs, rather than full sentences. The experimenter said: "We're going to take it in turns to ask my dog for some cards. Are you ready? I'll go first. 'The bird who's happy?' [Dog answers 'Yes' or 'No']. So now it's your go. Ask the dog for the bird who's sad. Just say 'The bird who's sad' [Dog answers 'Yes' or 'No']." When the dog answered 'Yes', the experimenter took the card depicting the relevant animal from the dog's box of cards, and placed it on – as appropriate – her own bingo grid or, with the child's help, the child's bingo grid. The dog's responses were structured such that the child always won the bingo game (i.e., was the first player to complete the bingo grid) on Days 1, 3, and 5, and the

experimenter on Days 2 and 4, resulting in an overall win for the child. In order to provide pragmatic motivation for the use of complex noun phrases (e.g., *the bird who's sad*, as opposed to simply *the bird*), the bingo grid contained two of each animal, with opposite properties (e.g., *the bird who's happy* vs. *the bird who's sad*; *the chicken who's big* vs. *the chicken who's small*), requested on subsequent turns by the child and the experimenter (as shown in Table 1). Two different versions of the game were created, with different pairings of animals and adjectives, the first used on Days 1, 3, and 5, the second on Days 2 and 4. The allocation of NPs to the experimenter versus the child, and the order of the trials was varied within each version, but was not subject to any between-subjects variation: within a particular group (Experimental/Control) all children had identical training.

Children in the Control group received similar training to the Experimental group, except that instead of complex NPs (e.g., *the bird who's happy*), they heard and repeated semantically matched simple adjectival NPs (e.g., *the happy bird*), as shown in the two rightmost columns of Table 1.

2.4.2. *Simple-question training*

The aim of this part of the session was to train children on simple questions (e.g., *Is the bird cleaning?*), resulting in the formation of a simple question schema (**Is [THING] [ACTION]ing?**) that children in the Experimental group – but crucially not the Control group – could combine with the trained complex-noun-phrase schema (**the [THING] who's [PROPERTY]**) to yield a complex-question schema (**Is [the [THING] who's [PROPERTY]] ACTIONing?**). Simple question training was identical for the Experimental and Control groups, and took the form of a game in which the child repeated questions spoken by the experimenter (see Table 2), subsequently answered by the same talking dog toy from the NP training part of the session.

The experimenter first explained that: "We are going to ask the dog some questions. We'll see an animal on the card and try to guess what the animal is doing on the other side of the card." On each trial, the experimenter first showed the face of the card depicting the animal doing nothing in particular and said, for example: "On this one, here's a bird. I wonder if the bird is cleaning. Let's ask the dog. Copy me. Is the bird cleaning." After the child had attempted to repeat the question, the dog responded (e.g., "No, he's having his dinner"), and the experimenter turned the card to show an illustration depicting the answer. As for the NP training, two different versions of the game were created, with different pairings of animals and actions, the first used on Days 1, 3, and 5, the second on Days 2 and 4, with the order of presentation varied within each version. All children, regardless of group, had identical simple-question training.

TABLE 2. *Day 1 simple-question training for all children*

Experimenter says, child repeats …	Talking dog toy answers
Is the bird cleaning?	No, he's having his dinner
Is the fish swimming?	Yes, he's swimming in the pond
Is the whale falling?	No, he's OK!
Is the frog clapping?	No, he's croaking
Is the chicken crying?	Yes, he's a bit sad today
Is the cow drinking?	Yes, he's drinking some water
Is the bear drawing?	No, he's having his breakfast
Is the horse painting?	No, he's going for a run
Is the dog walking?	No, he's fetching his ball
Is the cat jumping?	Yes, he's jumping up and down
Is the mouse laughing?	Yes, he heard a funny joke
Is the rabbit hopping?	Yes, and he's having lots of fun

Note that, in order to encourage schema combination, an identical set of animals featured in the NP training (e.g., *the bird who's sad*) and simple-question training (e.g., *is the bird cleaning?*). This overlap is not strictly speaking necessary according to a schema-combination account, which assumes that children are capable of combining FULLY ABSTRACT schemas (here, **Is [THING] [ACTION]ing? + the [THING] who's [PROPERTY] → Is [the [THING] who's [PROPERTY]] ACTIONing?**). However, we considered it likely that lexical overlap would be helpful, given the evidence from syntactic priming studies that such overlap highlights structural similarity (the so-called 'lexical boost'; e.g., Rowland et al., 2012), and – more generally – evidence that learners retain highly detailed representations of individual exemplars (Ambridge, 2019).

2.5. TEST PHASE: COMPLEX QUESTIONS

The aim of the test phase was to investigate children's ability to produce complex questions (e.g., *Is the crocodile who's hot eating?*) by combining trained complex-NP and simple-question schemas **(Is [the [THING] who's [PROPERTY]] ACTIONing?**). Because we were interested in training an abstract schema, rather than particular lexical strings, the target complex questions for the test phase used only animals, verbs, and adjectives that were not featured during training (see Table 3).

The game was very similar to that used in the simple-question training, except that children were told: "This time you are not going to copy me. I will tell you what to ask, and you can ask the dog." For each trial, the experimenter held up the relevant card and said (for example): "Two crocodiles: hot and cold [points to each crocodile; one wearing swimwear on a beach; the other wearing winter clothing in snow]. I wonder if the crocodile who's hot is eating. Ask the dog if the

TABLE 3. *Complex-question test session (Day 5) for all children*

Child's target complex question	Talking dog toy answers
Is the crocodile who's hot eating?	Yes, he's having his breakfast
Is the penguin who's cold dancing?	No, he's skating
Is the elephant who's thin hiding?	Yes, he doesn't want anyone to find him
Is the giraffe who's fat driving?	Yes, he's driving in his car
Is the goat who's tall singing?	No, he's not very good at singing
Is the hedgehog who's short playing?	No, he's having his dinner
Is the lion who's clean running?	Yes, he's really fast
Is the monkey who's dirty climbing?	Yes, he's going really high up in the tree
Is the panda who's heavy cooking?	Yes, he's making his tea
Is the tiger who's light sitting?	No, he's going for a walk
Is the zebra who's fast waving?	No, he's trying to catch a fly
Is the duck who's slow flying?	No, he's saying "quack quack"

crocodile who's hot is eating." Note that this prompt precludes the possibility of the child producing a well-formed question simply by repeating part of the experimenter's utterance. As before, the dog then answered (e.g., "Yes, he's having his breakfast"), and the experimenter turned the card to show the relevant animation. Each child completed 12 test trials in random order.

2.6. EXCLUSION CRITERIA

In order to ensure that both the Experimental and Control groups were made up of children who had successfully completed the training, we followed our pre-registered exclusion criteria, which specified that "any child who does not correctly repeat at least half of the noun phrases and at least half of the questions on all five days will be excluded … All children who complete the training and test to criterion (outlined above) will be included, and any who do not will be replaced". On this criterion, we excluded 21 children.

2.7. CODING

All participants produced scorable responses for all trials, with no missing data (i.e., all responses were clearly some attempt at the target question). Presumably this was due to our extensive training and strict exclusion criteria which ensured that children were competent and confident in putting questions to the talking dog in response to prompts from the experimenter. Responses were coded according to the scheme shown in Table 4, which also shows the number of responses in each category, for each group (allowing *that*-for-*who* substitutions; see below).

Unfortunately, we failed to anticipate the possibility of *that*-for-*who* substitutions in our pre-registration, and so did not specify in advance whether or not we would score such utterances as correct. We therefore used both a strict coding scheme, in which production of *that* in place of *who* was scored as an

error, and a lenient coding scheme, in which such substitutions were ignored (i.e., any such utterance was scored as a correct target question, provided that it contained no other errors). We concur with an anonymous reviewer who suggested that "questions with this 'error' should definitely be coded as correct—the error does not change the structure of the question. So … the appropriate result to emphasize … is the one using the 'lenient' scoring criteria". For this reason, the lenient coding scheme is the one used in both our summary data (see Table 4) and our headline finding. However, as noted by a second anonymous reviewer, this coding decision is not theory-neutral,

TABLE 4. *Coding scheme and counts of each response*

Classification	Description	Example	*N* Exp	*N* Ctrl
Correct	Exact production of the target question	*Is the crocodile who's hot eating?*	308	209
	or with production of *that* in place of *who*	*Is the crocodile that's hot eating?*	16	83
Repetition	Any utterance starting "I wonder"	*I wonder if the crocodile who/that's hot is eating?*	2	11
Aux-doubling	Otherwise well-formed question with 'doubled' auxiliary	*Is the crocodile who/that's hot is eating?*	101	209
Resumption	Final verb is preceded by additional auxiliary +subject	*Is the crocodile who/that's hot, is she/he/it eating?*	18	15
Statement	Complex utterance but phrased as a statement/ intonation question rather than a syntactic *yes/no* question	*The crocodile who/that's hot is eating?*	69	21
Structure dependence*	Question with 'unmoved' auxiliary	*Is the crocodile who/that hot is eating?*	3	1
Simple	Simple question with adjectival NP	*Is the hot crocodile eating?*	3	11
Miscellaneous**	Substitution/omission of animal, verb, or adjective; multiple or other errors	*The crocodile that hot will be eating? / Is the crocodile eating? / Is the cold crocodile running?*	212	179

NOTES: As detailed in the main text, these counts reflect the use of a lenient version of the coding scheme which allows (i.e., ignores) production of *that* in place of *who*.

*'Structure dependence' errors (as they are termed in Ambridge et al., 2008), merit a category of their own, as they are the errors that children famously do not make (Chomsky, 1980; Crain & Nakayama, 1987; though see Ambridge et al., 2008), but would be expected to make if they had internalised a non-structure-dependent movement rule 'move the first auxiliary').

**The most common type of miscellaneous error involved omission of the adjective altogether (e.g., *Is the crocodile eating?; Is he eating?*). For further examples, readers are invited to inspect the raw data, available for download https://osf.io/e2q54/.

and assumes that "learners consider the schema 'the [THING] who's [PROPERTY]' interchangeable with the schema 'the [THING] that's [PROPERTY]'". In our view, this assumption is justifiable, on the basis that children presumably have considerable evidence from the input that these two schemas have extremely similar distributions and are thus largely (if not entirely) interchangeable. Furthermore, only the lenient coding scheme is fair to generativist–nativist accounts, which assume that both *that* and *who* are members of the same functional category, and so can be used interchangeably (in semantically appropriate contexts). (A reviewer suggested that "generativist–nativist proposals would predict no effect of the training regardless of whether responses were coded with strict or lenient criteria". Our view is that generativist–nativist proposals would predict no effect under the lenient criteria, but could explain away any (apparent) effect observed under the strict criteria as the child simply learning that the experimenter seemed to prefer the use of *who* than *that* in the context of the experiment.)

In order to check reliability, all responses were independently coded by two coders: the first and final authors. At the first pass, the coders showed 100% agreement with regard to the classification of responses as correct (1) or erroneous (0), with the only disagreements relating to the classification of error types (84 cases for an overall agreement rate of 94.3%). All of these discrepancies related to ambiguities in the coding scheme and, following discussion, were eliminated for 100% agreement.

2.8. PREDICTIONS

Our two pre-registered predictions were as follows:

1. The Experimental group will produce significantly more correctly formed complex questions (out of a maximum of 12) than the Control group (as determined by maximal mixed-effects models; e.g., Barr, Levy, Scheepers, & Tily, 2013).
2. Significantly more children in the Experimental than Control group will produce at least one correctly formed question (as determined by chi-square test).

The reason for including this second prediction, despite the fact that, as a categorical statistic, the chi-square test has much lower power, was that it taps more directly into the question of whether our training regime is sufficient to 'create' the ability to produce a complex *yes/no* question, as opposed to simply boosting this ability.

2.9. ANALYSES

The data were analysed according to our pre-registered analysis plan. To compare the number of correct questions produced by each group, we ran a linear mixed-effects regression model in R (R Core Team, 2017) with random intercepts for participant and item, using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). No random slopes were included because, as per the analysis plan, we simplified the model using the procedure outlined by Barr et al. (2013) in order to enable convergence. In general, we were able to include all the fixed effects specified in the analysis plan: Group (Experimental vs Control; coded as 1 vs. 0), Age in months, Score on the standardised grammar test (CELF Word Structure), Days taken to complete the five training blocks, Number of noun phrases correctly repeated during training, and Number of simple questions correctly repeated during training. However, in some cases, one or more nonsignificant control predictors had to be removed in order to enable the model to converge (as set out in detail below). *P*-values were obtained using the model-comparison (likelihood ratio test) procedure. To compare the number of participants in the Experimental and Control groups producing at least one correct complex question, we ran a 2×2 chi square analysis: Group (Experimental/Control) × children producing / not producing at least one correctly formed complex question. We also ran a number of exploratory non-pre-registered analyses, which are clearly differentiated below from the planned pre-registered analyses.

3. Results

Table 5 shows the outcome of the linear mixed-effects regression analysis. In order to enable the model to converge, we had to remove the control predictors for the number of (a) noun phrases and (b) simple questions correctly repeated

TABLE 5. *Model and model comparisons using strict coding scheme*

	Model				Model comparisons		
Fixed effect	Estimate	SE	*z*-value	*p* (z)	AIC	ChiSq	*p* (ChiSq)
(Intercept)	–14.62	3.86	–3.79	.000	1183.50		
Group = Exp (vs. Control)	1.51	0.61	2.48	.013	1187.60	6.09	.014
Age (months)	0.07	0.04	1.57	.116	1183.90	2.47	.116
Test period (days)	–0.16	0.20	–0.81	.419	1182.10	0.66	.418
Word structure test	0.43	0.12	3.51	.000	1194.40	12.91	.000

TABLE 6. Model and model comparisons using lenient coding scheme (allows *that* for *who* substitution)

	Model				Model comparisons		
Fixed effect	Estimate	SE	*z*-value	*p* (z)	AIC	ChiSq	*p* (ChiSq)
(Intercept)	–17.50	3.54	–4.94	.000	1228.50		
Group = Exp (vs. Control)	0.72	0.54	1.33	.183	1228.20	1.77	.184
Age (months)	0.10	0.04	2.40	.016	1232.20	5.78	.016
Test period (days)	–0.01	0.18	–0.04	.972	1226.50	0.00	.972
Word structure test	0.50	0.11	4.41	.000	1247.00	20.50	.000

during training. As predicted, the Experimental group significantly outperformed the Control group on the proportion of correct questions produced ($M = 0.42$, $SD = 0.49$ vs. $M = 0.29$, $SD = 0.45$, chi-square = 6.09, $p = .014$). However, this finding is contingent upon a very strict interpretation of the 'Correct' question criterion, which requires children to produce exactly the target question, with no substitutions (e.g., *Is the crocodile who's hot eating?*). If this criterion is relaxed to allow substitution of relativiser *that* for *who* (e.g., *Is the crocodile that's hot eating?*; 99 instances across the two groups), no significant difference between the Experimental and Control groups is seen ($M = 0.44$, $SD = 0.50$ vs. $M = 0.40$, $SD = 0.49$, chi-square = 1.77, $p = .18$, n.s; see Table 6). A significant positive effect of age in months is observed using the lenient, but not strict, coding scheme. Although we did not pre-register any predictions regarding age, we would fully expect older children to perform better at this task; hence this finding bolsters our conclusion that the lenient coding scheme better captures children's performance.

Because null results are difficult to interpret, particularly for studies with relatively low power, we additionally ran an exploratory (i.e., not pre-registered) Bayesian mixed-effects model with a wide prior for all predictor variables ($M = 0$, $SD = 2$), using brms (Bürkner, 2017). This model was run only for data coded using the lenient coding scheme. Because Bayesian models are more robust to convergence failure, we were able to build a maximally conservative model with all control predictors, and by-item random slopes for group, age, score on the CELF Word Structure test, days taken to complete the training, and the number of (a) simple questions and (b) noun phrases correctly repeated during training. This model is shown in Table 7. Although the credible interval for the effect of group (Experimental vs. Control) includes zero ($M = 0.71$ [–0.44, 1.83]), we do not see a distribution that would constitute positive evidence for the ABSENCE of this effect; i.e., a narrow credible

TABLE 7. *Bayesian model using lenient coding scheme*

Fixed effect	Estimate	SE	95 CI Lower	95 CI Upper	Eff Samples	Rhat	*p* (MCMC)
(Intercept)	–14.46	4.93	–24.19	–4.78	2026.00	1.00	.002
Group = Exp (vs. Control)	0.71	0.58	–0.44	1.83	1147.00	1.00	.105
Age (months)	0.10	0.04	0.01	0.19	966.00	1.01	.015
Test period (days)	–0.03	0.21	–0.42	0.38	1055.00	1.00	.437
Word structure CELF	0.53	0.13	0.28	0.80	1073.00	1.00	0
Training NPs repeated	–0.27	0.21	–0.67	0.15	1151.00	1.00	.101
Training Qs repeated	–0.49	0.58	–1.66	0.61	6112.00	1.00	.202

interval centred around zero. Indeed, the *pMCMC* value of 0.105 indicates a 90% probability of an effect of group greater than zero. Thus, we cannot rule out the possibility that a between-groups difference may have been observed on this measure, had we been able to achieve greater power through the use of a larger sample.

However, given that the aim of the study was to train a complex question structure, potentially more informative (if less sensitive) is the chi-square analysis of children who produced / failed to produce at least one complex question. This analysis found no difference between groups using either the strict coding criteria (Experimental group 40/61 vs. Control group 35/61; chi-square = 0.55, $p = .46$, n.s.) or the lenient criteria (Experimental group 41/61 vs. Control group 42/61; chi-square = 0, $p = 1$, n.s.). The fact that, even with no training (i.e., in the Control group), around two-thirds of children were able to produce at least one correctly formed question suggests one possible reason for our failure to observe an effect of our experimental manipulation on the lenient criteria: The training was superfluous because most children of this age and from this population already have the ability to form complex questions, at least with *that*, if not with *who*.

3.1. EXPLORATORY NON-PRE-REGISTERED ANALYSES

In order to explore the possibility that, as a whole, our participant group was too old and/or too advanced to have not yet acquired this ability, we re-ran the analyses above, looking only at (a) children below the mean age (67 months; $N = 66$) and (b) children scoring below the mean on the CELF Word Structure test (19.69, $N = 66$). These analyses, like all reported in the current section, were not pre-registered, and were conceived after having seen the data. Both of these subgroup analyses yielded almost identical findings to the main, all-participants analyses, and so are not reported in detail.

TABLE 8. *Model for analysis of first test trial only, using lenient coding scheme (allows that for who substitution)*

Effect	Estimate	SE	*z*-value	*p* (z)
(Intercept)	–2.97	7.53	–0.40	.693
Group = Exp (vs. Control)	**1.14**	**0.48**	**2.38**	**.018**
Age (months)	0.07	0.03	2.14	.033
Test period (days)	0.25	0.15	1.62	.105
Word structure test	0.39	0.12	3.29	.001
Training NPs repeated	0.10	0.51	0.20	.839
Training Qs repeated	–2.63	1.33	–1.99	.047

Another possibility, however, is that at least some children somehow learned to produce complex questions during the test session. In order to explore this possibility, we re-ran the lenient-coding analysis above, looking only at the first test trial completed by each child (two children were excluded because trial-order was not recorded). This analysis required the use of a standard (non-mixed-effects) binomial linear regression model, since it includes only one trial per participant. Perhaps surprisingly, this analysis did suggest some evidence for an effect of group (see Table 8): A significantly higher proportion ($p = .018$) of children in the Experimental group ($M = 0.42$, $SD = 0.50$) than the Control group ($M = 0.27$, $SD = 0.45$) produced a correctly formed question (with either *who* or *that*) on their first attempt (25/60 and 16/60 children, respectively).

Repeating the same analysis using a chi-squared test did not yield a significant effect ($p = .12$, n.s.), suggesting that the effect observed in the linear model is contingent on partialling out the potentially confounding effects of the control predictors (particularly age, score on the word-structure test, and number of training questions repeated, all of which were significant). Indeed, removing all control predictors from the binomial model resulted in an effect of group that was no longer significant (*Estimate* = 0.68, $SE = 0.39$, $z = 1.72$, $p = .08$).

A final non-pre-registered analysis conceived after having seen the data compared the groups on the number of auxiliary-doubling errors (*Is the crocodile who's/that's hot is eating?*), which are particularly common for complex questions (Crain & Nakayama, 1987; Ambridge et al., 2008). This analysis was not theoretically motivated, and arose purely from the observation (see Table 4) that, numerically, children in the Control group produced considerably more such errors than children in the Experimental group. In order to enable the model to converge, it was necessary to remove the non-significant control predictors that explained the least variance: age, days taken to complete training, and score on the CELF word structure test. This model (see Table 9) revealed a significant main effect of group (chi-square = 6.61, $p = .01$), such that children in the Control group ($M = 0.28$, $SD = 0.44$) produced approximately

TABLE 9. *Model and model comparisons for number of auxiliary-doubling errors*

					Model comparisons		
Effect	Estimate	SE	*z*-value	*p* (z)	AIC	ChiSq	*p* (ChiSq)
(Intercept)	–13.00	5.33	–2.44	.015	937.54		
Group = Exp (vs. Control)	–1.80	0.70	–2.56	.010	942.15	6.61	.010
Training NPs repeated	0.46	0.23	1.96	.050	939.56	4.02	.045
Training Qs repeated	1.67	1.05	1.60	.110	938.81	3.26	.071

double the proportion of auxiliary-doubling errors compared with children in the Experimental group ($M = 0.13$, $SD = 0.34$), again suggesting some effect of the training regime.

4. Discussion

A central question in language acquisition is how children master sentence types that they have seldom, if ever, heard. The aim of the present study was to test the prediction that, for one such sentence type, complex questions (e.g., *Is the crocodile who's hot eating?*), children could combine schemas learned, on the basis of the input, for complex noun phrases (**the [THING] who's [PROPERTY]**) and simple questions (**Is [THING] [ACTION]ing?**) to yield a complex-question schema (**Is [the [THING] who's [PROPERTY]] ACTIONing?**). To investigate this possibility, 122 children aged four to six years were trained on simple questions (e.g., *Is the bird cleaning?*) and either (Experimental group) complex noun phrases (e.g., *the bird who's sad*) or (Control group) matched simple noun phrases (e.g., *the sad bird*). In fact, on most measures, the two groups did not differ on their ability to produce novel complex questions at test.

We can see a number of possible reasons for the failure to observe a training effect in the lenient-coding version of the pre-registered analyses. The first, of course, is that there is no effect to find, and that children do not learn to form complex questions by combining schemas in the manner proposed. The second is that our study was under-powered in terms of the number of participants. The supplementary Bayesian analysis is consistent with this possibility with regard to the continuous measure (number of correct complex questions per group), even on the lenient scoring criteria that allow *that* for *who* substitution. However, inconsistent with this possibility, the categorical (chi-square) analysis found that (non-significantly) FEWER children in the Experimental than Control group (41 vs. 42) produced at least one correct complex

question at test. The third possibility is that the majority of children tested already had the linguistic knowledge and ability to produce correctly formed complex questions, meaning that the intervention had little effect. Consistent with this possibility, over two-thirds of children in the Control group were able to produce at least one complex question, despite not receiving the (by hypothesis) crucial element of training (complex noun phrases).

On the other hand, the findings of two non-pre-registered analyses – which should therefore be treated with the utmost caution – suggest that the training regime did have at least some effect on children's linguistic productions. First, children in the Experimental group produced significantly fewer auxiliary-doubling errors (*Is the crocodile who's/that's hot is eating?*) than children in the Control group; an error that is often taken as being particularly diagnostic of failure to master complex *yes/no* questions (e.g., Crain & Nakayama, 1987; Ambridge et al., 2008). Of course, as we have already seen, this did not translate into a higher rate of correct questions on the part of the Experimental group (at least under the lenient coding scheme). Rather, children in the Experimental group seemed to avoid auxiliary-doubling errors by producing more statements / intonation questions (we did not attempt to differentiate the two) such as *The crocodile who's/that's hot is eating?*, and more miscellaneous utterances (e.g., *Is the crocodile eating?*). A possibility, then – albeit a very tentative and speculative one – is that children in the Experimental group had learned, in the context of complex questions, to inhibit the tendency to produce an auxiliary (or agreeing verb form) after a noun phrase. This tendency is a very pervasive one, since – given the *n*-gram statistics of English – such a sequence is found in a high proportion of (complete, non-interrogative) utterances (e.g., Lewis & Elman, 2001; Reali & Christiansen, 2005; Ambridge et al., 2008). On this account, children in the Experimental group who had not yet learned how to form complex questions had at least learned what NOT to do and so – compared with the control group – were more likely to produce statements or intonation questions as an evasion strategy.

The second non-pre-registered finding was that a significantly greater proportion of children in the Experimental than Control group were able to produce a complex question (even under the lenient coding scheme that allows either *who* or *that*) on their first attempt at test. This raises the intriguing possibility that our training regime was successful in teaching – or at least reinforcing – a complex question construction in the Experimental but not Control group, but that this group difference had washed out by the end of the test session; i.e., that the test session itself constituted some kind of training on complex-question formation. How plausible is this? If our participants had excellent knowledge of the relationship between indirect and direct questions (e.g., *I wonder if the crocodile is eating* → *Is the crocodile eating?*), then the test-session prompts in fact constitute clear evidence for the correct structure of

complex questions (e.g., *I wonder if the crocodile [who's hot] is eating* → *Is the crocodile [who's hot] eating?*). Perhaps a more plausible possibility is that knowledge of complex-question formation is not all or none, but graded. Our participants (or some of them) started out with a weak representation of this construction, which was boosted by both (a) our training regime, hence the Experimental > Control group difference on the first test trial, and (b) repeated practice at producing complex questions, hence the finding that this difference had washed out by the end of the test session. But, to emphasise, since this analysis was not pre-registered, it contains only extremely preliminary evidence of an effect of our training manipulation. This finding speaks to the broader issue of when experimental studies should focus on training new knowledge, and when they should focus on priming or facilitation of already acquired knowledge, an issue to which we return shortly.

Returning to our pre-registered analyses, one difference between groups did emerge: the Experimental group produced significantly more questions than the Control group who used the PARTICULAR complex noun phrase that was trained; i.e., **the [THING] *who's* [PROPERTY]**, as opposed to **the [THING] *that's* [PROPERTY]**. That is, we observed a significant between-groups difference according to strict coding criteria that require the use of relativiser *who*, which the Experimental group heard during complex noun-phrase training. We can see three possible explanations for this finding.

The first is that this is simply a task effect: irrespective of training, the majority of children were already relatively adept at producing complex questions with both *that's* and *who's,* but – all else being equal – preferred to use the former, perhaps due to a frequency effect (*that's* is around five times more frequent than *who's* in Google search results). On this account, children in the Experimental group learned nothing more than that the experimenter seemed keen for them to use *who's* rather than *that's*, in contexts in which either is permissible.

The second possible explanation is that our manipulation was in fact successful in teaching the particular complex-question structure (**Is [the [THING] who's [PROPERTY]] ACTIONing?**). On this view, the majority of children began the study with a different complex-question structure based around *that's* (**Is [the [THING] that's [PROPERTY]] ACTIONing?**), which children in the Control group used successfully at test. Children in the Experimental group, however, were taught a NEW complex-question structure based around *who's*, and frequently used this structure at test. This possibility is consistent with a constructivist account of language acquisition in which children's syntactic structures are initially built around particular lexical items, here *who's* and *that's* (e.g., Tomasello, 2003). It is also consistent with an exemplar account under which individual lexical strings are never replaced by more abstract representations, but are retained and form the basis

for subsequent productions, via either direct reuse or on-the-fly similarity-based analogy (e.g., Chandler, 2010; Ambridge, 2019). Indeed, because the nouns, verbs, and adjectives differed at training and test, participants cannot have been directly reusing lexical strings from training; some kind of generalisation must have occurred.

The third possible explanation is that children already had abstract knowledge of complex-question formation at the start of the study, with our training giving children in the Experimental group the tools to apply this knowledge to a relatively unfamiliar relativiser. For example, children might not have known before the study that *who* functions as a relativiser. Alternatively, they might have known in principle that both *that* and *who* are relativisers, but had very little experience in using the latter in production. This possibility is consistent with a generativist–nativist account, under which syntactic knowledge crucial for complex question formation (e.g., movement, structure dependence) is present at the start of language acquisition (e.g., Chomsky, 1980; Crain & Nakyama, 1987).

The present findings do not allow us to mediate between these three possibilities. We therefore end by suggesting a number of possible future studies that may be able to do so and to provide a stronger test of the claim that complex questions are learned via schema-combination.

One possibility would be to simply adapt the current design in one or more ways. For example, since, all other things being equal, children seem to prefer to use *that's* than *who's* in complex NPs, switching entirely to *that's* in a new study may allow for the observation of between-group differences that may have been masked, in the present study, by the experimenter's apparent insistence, during training, that children in the Experimental group use a dispreferred relativiser. The use of this more frequent relativiser during training may also allow for the testing of younger children. It may be possible to find a sweet-spot age at which most children cannot yet produce complex questions (which could be determined by a screening task), but can be trained to do so by something like the present regime.

However, a new study along these would face two perhaps insurmountable difficulties. The first problem is one of resources: the present version required around 18 months of full-time testing (122 children plus 21 drop-outs, each tested for five days), and versions with younger children – who generally show higher drop-out rates – would presumably take longer still. The second (related) problem is one of power. The present Bayesian analysis suggests that the current study was under-powered, and power is a function of both sample size and effect size. Given the resource-heavy nature of this type of training study, substantially increasing the sample size is infeasible. But increasing the effect size is probably even less feasible. As two anonymous reviewers noted, given the amount of linguistic input that children receive every day, any

training that we provide is a drop in the ocean, and is all but guaranteed to have a very small effect. On this view, the surprising aspect of the present study is not the failure to observe a between-groups difference on our primary dependent measures, but the fact that any differences (e.g., increased use of *who* vs. *that* in the Experimental vs. Control group) were observed at all.

It may well be, then, that the present design – even with some modifications – is not well suited to testing the schema-combination account of children's acquisition of complex *yes/no* questions (or other complex constructions), and that a different method altogether is required. One possibility (suggested by an anonymous reviewer) would be to use an artificial language learning task. The obvious advantage of such a task is that, by definition, children have no knowledge of the artificial language at the start of the task, which ensures that any learning – and any observed between-groups difference – can only plausibly be attributed to the training regime. A disadvantage, however, is that, in order to test children's ability to combine two newly learned schemas, the artificial language would have to be considerably more complex than those used in previous child studies (or at least those with real-world semantics; e.g., Hudson Kam & Newport, 2005; Hunter & Lidz, 2013; Culbertson & Newport, 2015; Hudson Kam, 2015; Shuler, Yang, & Newport, 2016; Culbertson, Jarvinen, Haggerty, & Smith, 2018; Brown, Smith, Samara, & Wonnacott, in press). For example, in Brown et al. (in press), although the semantic constraint present in the language was relatively simple (animals vs. vehicles), and the children were relatively old (six years), it was learned only when it was exceptionless, and only by children who were able to verbalise it explicitly.

Given the difficulties with both familiar- and artificial-langue training studies, it may be that a third approach is required. One possibility (suggested by a second anonymous reviewer) is to abandon the very difficult task of creating new knowledge, and instead to focus on priming or facilitating already-acquired knowledge. Such an approach would treat the fact that children have already acquired some knowledge of (in this case) complex questions not as a bug in the experimental design, but as a feature. Many syntactic priming studies (see Mahowald, James, Futrell, & Gibson, 2016; Branigan & Pickering, 2017, for reviews) have found that, having heard and/or produced a particular syntactic construction (e.g., the passive, as in *The vase was broken by the hammer*), both adults and children show an increased tendency to use this construction when subsequently asked to describe a picture or animation (e.g., to produce *The bricks were pushed by the digger*, rather than *The digger pushed the bricks*). There is also some evidence (summarised in Mahowald *et al.*'s, 2016, meta-analysis) that this priming effect is increased when lexical material is shared between the prime and the target; the so-called LEXICAL BOOST.

This phenomenon potentially allows for a test of the schema-combination account of complex *yes/no* questions outlined here. Both this account and accounts that posit more abstract knowledge, including generativist-nativist accounts, would predict that a complex NP (e.g., *a bird who's sad)* should constitute a better prime than a semantically matched simple NP (e.g., *a sad bird*), for a complex *yes/no* question that shares little to no lexical overlap with either (e.g., *Is the crocodile that's hot eating?*). However, only the schema-combination account predicts a LEXICAL BOOST: that the priming effect will increase as a function of the lexical overlap between the prime and the target. For example, given the target *Is* ***the*** *crocodile* ***that's*** *hot eating?*, a greater priming effect would be predicted following ***the*** *bird* ***that's*** *sad* than *a bird who's sad*. This prediction follows from the schema-combination account on the assumption that, when more abstract schemas are formed, the concrete utterances that gave rise to them are not discarded, but remain in memory (e.g., Abbot-Smith & Tomasello, 2006). Indeed, some versions of this account (see Ambridge, 2019, for a review) assume that so-called 'abstract schemas' are not represented independently at all, and reflect nothing more than clusters of stored exemplars. Of course, generativist-nativist accounts do not ACTIVELY RULE OUT this type of lexical boost. However, this effect could be explained only by positing some supplementary mechanism, above and beyond the fully abstract rules used to form complex questions. Potentially, then, syntactic priming might constitute a far more powerful and less resource-intensive way of testing the schema-combination hypothesis in the future.

In the meantime, the question of how children learn to produce complex *yes/no* questions remains unanswered by the present training study. Although children in the Experimental group did show some evidence of generalising complex NPs learned during training into complex-question production at test (particularly with unplanned exploratory analyses), the present data do not allow us to mediate between generativist, constructivist, and task-based accounts of this finding, leaving open the possibility of some role for innate knowledge. Our hope is that, nevertheless, the present paper – and in particular what we have learned about the strengths and weaknesses of different design decisions – will inspire future work that will indeed be able to mediate between these competing theoretical accounts of this phenomenon, both in its own right, and as a test case for children's language acquisition more generally.

REFERENCES

Abbot-Smith, K. & Behrens, H. (2006). How known constructions influence the acquisition of other constructions: the German passive and future constructions. *Cognitive Science* **30**, 995–1026.

Abbot-Smith, K. & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review* **23**(3), 275–290.

Akhtar, N., Callanan, M., Pullum, G. K. & Scholz, B. C. (2004). Learning antecedents for anaphoric *one*. *Cognition* **93**(2), 141–145.

Ambridge, B. (2019). Against stored abstractions: a radical exemplar model of language acquisition. *First Language,* https://doi.org/10.1177/0142723719869731.

Ambridge, B. & Rowland, C.F. (2009). Predicting children's errors with negative questions: testing a schema-combination account. *Cognitive Linguistics* **20**(2), 225–266.

Ambridge, B., Rowland, C. F. & Pine, J. M. (2008). Is Structure Dependence an innate constraint? New experimental evidence from children's complex question production. *Cognitive Science* **32**(1), 222–255.

Ambridge, B., Rowland, C. F., Theakston, A. & Tomasello, M. (2006) Comparing different accounts of non-inversion errors in children's non-subject wh-questions: 'What experimental data can tell us?' *Journal of Child Language* **30**(3) 519–557.

Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* **68**(3), 255–278.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**(1), 1–48.

Berwick, R. C., Pietroski, P., Yankama, B. & Chomsky, N. (2011). Poverty of the stimulus revisited. *Cognitive Science* **35**(7), 1207–1242.

Branigan, H. P. & Pickering, M. J. (2017). An experimental approach to linguistic representation. *Behavioral and Brain Sciences* **40**, e282.

Brown, H., Smith, K., Samara, A. & Wonnacott, E. (in press). Semantic cues in language learning: an artificial language study with adult and child learners. https://doi.org/10.31234/osf.io/7hq2c.

Bürkner, P. (2017). brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**(1), 1–28.

Chandler, S. (2010). The English past tense: analogy redux. *Cognitive Linguistics* **21**, 371–417.

Chomsky, N. (1980). In M. Piatelli-Palmarini, *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Clark, A. & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. Oxford: Wiley-Blackwell.

Cohen, J. (1992). A power primer. *Psychological Bulletin* **112**(1), 155–159.

Cowie, F. (1998). *What's within? Nativism reconsidered*. New York: Oxford University Press.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences* **14**(4), 597–612.

Crain, S. & Nakayama, M. (1987). Structure dependence in grammar formation. *Language* **63**, 522–543.

Crain, S. & Pietroski, P. (2001). Nature, nurture and universal grammar. *Linguistics and Philosophy* **24**, 139–186.

Crain, S. & Thornton, R. (1998). *Investigations in universal grammar*. Cambridge, MA: MIT Press.

Culbertson, J., Jarviene, H., Haggerty, F. & Smith, K. (2018). Children's sensitivity to phonological and semantic cues during noun class learning: evidence for a phonological bias. *Language* **95**(2). 268–293.

Culbertson, J. & Newport, E. L. (2015). Harmonic biases in child learning: in support of language universals. *Cognition* **139**, 71–82.

Dąbrowska, E. (2000). From formula to schema: the acquisition of English questions. *Cognitive Linguistics* **11**(1/2), 83–102.

Dąbrowska, E. & Lieven, E. V. M. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* **16**(3), 437–474.

Fitz, H. & Chang, F. (2017). Meaningful questions: the acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition* **166**, 225–250.

Fodor, F. D. & Crowther, C. (2002). Understanding stimulus poverty arguments. *Linguistic Review* **19**, 105–145.

Getz, H. R. (2019). Acquiring *wanna*: beyond Universal Grammar. *Language Acquisition* **26**(2), 119–143.
Goldberg, A. E. & Michaelis, L. A. (2017). One among many: anaphoric one and its relationship with numeral one. *Cognitive science* **41**, 233–258.
Hudson Kam, C. L. (2015). The impact of conditioning variables on the acquisition of variation in adult and child learners. *Linguistic Society of America* **91**(4), 906–937.
Hudson Kam, C. L. & Newport, E. L. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development*. **1**(2) 151-195.
Hunter, T. & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, **30**, 315-334.
Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D. & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive Science* **32**, 771–787.
Laurence, S. & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science* **52**, 217–276.
Legate, J. A. & Yang, C. (2002). Empirical re-assessment of stimulus poverty arguments. *Linguistic Review* **19**, 151–162.
Lewis, J. D. & Elman, J. L. (2001). Learnability and the statistical structure of language: poverty of stimulus arguments revisited. In B. Skarabela, S. Fish & A. H. J. Do (eds), *Proceedings of the twenty-sixth annual Boston University Conference on Language Development* (pp. 359–370). Somerville, MA: Cascadilla.
Lidz, J., Waxman, S. & Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. *Cognition* **89**(3), 295–303.
MacWhinney, B. (1975). Pragmatic patterns in child syntax. *Stanford Papers and Reports on Child Language Development* **10**, 153–165.
MacWhinney, B. (1987). The competition model. In B. MacWhinney (ed.), *Mechanisms of language acquisition* (pp. 249–308). Hillsdale, NJ: Erlbaum.
MacWhinney, B. (2004). A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* **31**(4), 883–914.
Mahowald, K., James, A., Futrell, R. & Gibson, E. (2016). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language* **91**, 5–27.
Pearl, L. S. & Mis, B. (2016). The role of indirect positive evidence in syntactic acquisition: a look at anaphoric *one*. *Language* **92**(1), 1–30.
Perfors, A., Tenenbaum, J. B. & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition* **118**(3), 306–338.
Peter, M., Chang, F., Pine, J. M., Blything, R. & Rowland, C. F. (2015). When and how do children develop knowledge of verb argument structure? Evidence from verb bias effects in a structural priming task. *Journal of Memory and Language* **81**, 1–15.
Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
Pullum, G. K. & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review* **18**(1/2), 9–50.
R Core Team (2017). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online <https://www.R-project.org/>.
Reali, F. & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: structure dependence and indirect statistical evidence. *Cognitive Science* **29**(6), 1007–1028.
Regier, T. & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition* **93**(2), 147–155.
Rowland, C. F. (2007). Explaining errors in children's questions. *Cognition* **104**(1), 106–134.
Rowland, C. F., Chang, F., Ambridge, B., Pine, J. M. & Lieven, E. V. (2012). The development of abstract syntax: evidence from structural priming and the lexical boost. *Cognition* **125**(1), 49–63.
Sakas, W. G. & Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition* **19** (2), 83–143.

Sampson, G. (1989). Language acquisition: Growth or learning? *Philosophical Papers* **18**, 203–240.
Scholz, B. C. & Pullum, G. K. (2002). Searching for arguments to support linguistic nativism. *Linguistic Review* **18**, 185–223.
Scholz, B. C. & Pullum, G. K. (2006). Irrational nativist exuberance. In R. Stainton (ed.), *Contemporary debates in cognitive science* (pp. 59–80). Oxford: Blackwell.
Shuler, K. D., Yang, C. & Newport, E. L. (2016). Testing the tolerance principle: children form productive rules when it is more computationally efficient to do so. Cog Sci 2016 Mindmodeling.org.
Stemmer, N. (1981). A note on empiricism and structure-dependence. *Journal of Child Language* **8**, 649–663.
Theakston, A. L., Lieven, E. V., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language* **28**(1), 127–152.
Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Wexler, K. (1998). Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. *Lingua* **106**(1/4), 23–79.
Wiig, E. H., Secord, W. & Semel, E. M. (2004). *CELF Preschool 2: clinical evaluation of language fundamentals preschool*. London: Pearson/PsychCorp.